# 联合注意力机制和多尺度特征的图像语义分割网络

张蕊[1)], 刘孟轩[2)*], 孟晓曼[1)], 武益超[1)]

[1)] (华北水利水电大学信息工程学院 郑州 450046)
[2)] (中国联合网络通信有限公司郑州市分公司 郑州 450052)
(liumx97@chinaunicom.cn)

**摘 要:** 针对卷积神经网络在图像语义分割时存在部分语义信息丢失、边界定位精度较低等问题, 提出联合注意力机制和多尺度特征的卷积神经网络. 首先基于注意力机制将网络提取到的多尺度特征进行加权融合, 然后采用扩张卷积和全局平均池化聚合多尺度目标信息, 最后采用边界精细粒度特征提取模块对分割边界进行优化. 在多尺度 PASCAL VOC2012 和高分辨率 Cityscapes 数据集上的实验结果表明, 所提网络的分割效果显著优于骨干网络 ResNet-101, 平均交并比分别提高 12.2 个百分点和 9.3 个百分点.

**关键词:** 语义分割; 注意力机制; 多尺度特征; 卷积神经网络
**中图法分类号:** TP391.41 **DOI:** 10.3724/SP.J.1089.2024.20018

# Research on Semantic Image Segmentation Network Combining Attention Mechanism and Multi-Scale Features

Zhang Rui[1)], Liu Mengxuan[2)*], Meng Xiaoman[1)], and Wu Yichao[1)]

[1)] (*College of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou* 450046)
[2)] (*China Unicom Zhengzhou Branch, Zhengzhou* 450052)

**Abstract:** To address the problems of partial semantic information loss and low accuracy of boundary localization when convolutional neural networks are used for image semantic segmentation, this paper constructs a convolutional neural network by combining the attention mechanism and multi-scale features. The model firstly combines the multi-scale features extracted by the network based on the attention mechanism for weighting, then uses dilated convolution and global average pooling to aggregate the multi-scale target information, and finally uses the boundary fine-grained feature extraction module to optimize the segmentation boundary. Experimental results on the multi-scale PASCAL VOC2012 and high-resolution Cityscapes datasets show that the segmentation effect of the network in this paper is significantly better than that of the backbone ResNet-101, and the average cross-merge ratio is improved by 12.2 percentage points and 9.3 percentage points, respectively.

**Key words:** semantic segmentation; attention mechanism; multi-scale features; convolutional neural network

图像语义分割是根据图像的视觉内容将每个像素点归类为其所属语义类别的技术[1], 不仅需要给每个目标对象进行语义标注, 而且需要精准地定位其在场景中的位置, 是计算机视觉的一个重

收稿日期: 2022-10-19; 修回日期: 2023-03-13. **基金项目**: 河南省科技攻关重点项目(192102210265, 202102210141). **张蕊**(1980—), 女, 博士, 教授, 硕士生导师, 主要研究方向为大数据、机器学习、激光点云数据处理、图像处理等; **刘孟轩**(1997—), 男, 硕士研究生, 论文通信作者, 主要研究方向为图像语义分割; **孟晓曼**(1998—), 女, 硕士研究生, 主要研究方向为图像处理、点云语义分割; **武益超**(1999—), 男, 硕士研究生, 主要研究方向为图像处理、激光点云数据处理.

要分支, 也是图像理解相关应用的核心技术. 图像语义分割在空间场景要素感知、城市精细化管理、智慧医疗等国家重大战略需求中发挥着举足轻重的作用. 作为场景感知的重要信息载体, 2D 图像凭借其结构化表达、格网分布, 且具有纹理信息等优势, 成为神经网络技术的重要数据来源; 同时, 大量可利用的、公开的 2D 图像数据集、数字成像设备的普及与其易操作性, 促使神经网络技术在 2D 图像上取得了诸多重大突破. 如何利用人工智能手段提升图像的解译能力, 实现自然地物目标的语义标识与边界精准特征提取, 成为亟待解决的难题.

传统的图像语义分割大多基于模型拟合或特征聚类[2-3], 但这些方法仅限于较为简单的地物目标, 依据目标对象的几何信息的语义信息表达能力较弱, 尤其是在复杂多变的城市场景中, 各地物目标存在因遮挡导致数据丢失、目标多样, 且尺度差异大、视觉透视导致同一目标近大远小等问题, 传统方法的分割结果会产生边界模糊、小目标语义信息易丢失等问题, 已经不能满足应用需求. 随着深度学习在处理具有规则结构的 2D 图像分割中取得了长足的进步[4], 涌现出一批经典的神经网络语义分割模型. 近年来, 计算机视觉领域的研究人员致力于在图像语义的属性表达中融入对象边界信息, 强化目标与目标之间的分割, 使得目标对象的特征提取更加精细, 从而提高语义分割的效果.

依据神经网络对图像特征图操作模式的不同, 目前基于深度学习的图像语义分割网络模型主要有 2 类.

(1) 基于卷积的神经网络模型. 其本质是以分层的方式自动识别特征, 通过将输入的数据矩阵与特征矩阵(卷积核)进行点乘加权求和, 提取原始图像中的特定局部特征. 卷积神经网络(convolutional neural network, CNN)最初是为了解决物体的平移问题而设计, 意味着该网络可以识别一个对象而不管其在图像中的位置如何, 即所谓的“平移不变性”. 但是, CNN 并不能很好地处理视点变化的其他效果, 如旋转和缩放; 此外, CNN 中的池化层会导致大量空间信息丢失, 忽略整体与局部之间的关联. 为此, 文献[5]通过上采样将提取的特征恢复到原始图像尺寸, 然后进行像素分类, 使网络可以接受不同尺度的输入; 但是, 池化操作在下采样过程中损失了大量空间几何信息, 导致不同语义类别之间边界模糊, 分割结果比较粗糙. 为了减少空间信息的丢失, U-Net[6]通过在上采样和与之对应的下采样之间构建跳跃连接进行特征融合, 形成一个对称的U型结构; SegNet[7]采用编码器-解码器模式, 其中, 编码器通过下采样获得图像高级语义信息并保存池化索引, 解码器在上采样时使用反池化操作, 利用池化索引填充图像空间信息并将特征图恢复到原始尺寸大小. U-Net 和 SegNet 均使用特征融合的方法恢复空间几何信息, 但不同的特征直接跨层融合会引入噪声, 损失语义分割精度. 文献[8]使用扩张卷积代替池化操作, 在参数量不变的情况下扩大感受野, 降低特征提取时由于特征图分辨率大幅减小造成的信息损失. 扩张卷积虽然可以有效地扩大感受野, 但同时导致网络获取小尺度目标信息的能力降低, 不利于小尺度目标的特征提取. 为此, Chen 等[9]提出 Deeplabv2, 包括扩张空间金字塔池化(atrous spatial pyramid pooling, ASPP), 通过并行使用多个不同扩张率的扩张卷积提取多尺度目标; Deeplabv3+[10]采用编码器-解码器结构, 将提取到的深层高级别语义特征与浅层低级别语义特征融合以获得更好的分割性能, 但由于仅融合了单个浅层特征, 未能充分利用各个尺度特征, 依然存在信息损失的情况. 针对多尺度目标提取问题, PSPNet[11]利用全局平均池化(global average pooling, GAP)聚合不同区域的上下文信息; RefineNet[12]利用远程残差连接使浅层的完善特征强化高级的语义特征; HRNet[13]通过并行卷积执行多尺度融合, 增强高分辨率表示. 但是, 这些方法在利用不同尺度的中间特征时, 仅简单地通过等权值拼接或求和的方式进行融合, 未能有效地挖掘并利用浅层特征和深层特征的互补信息, 这种简单的融合方式会损失不同尺度特征的特有信息.

(2) 基于注意力机制的神经网络模型. 在处理复杂视觉信息时能够抑制不重要的刺激, 将有限的神经计算资源分配给场景中的关键部分, 为更高层次的感知推理和更复杂的视觉处理任务提供更相关的信息. 该类模型的核心思想是通过一定手段获取每幅特征图重要性的差异, 并利用任务结果反向指导特征图的权重更新, 从而更高效地完成场景分割任务[14], 即基于注意力机制的网络模型的权重会随着全输入数据的变化而动态地调整, 这也正是其与 CNN 的本质区别. 例如, 采用单路注意力的 SE-Net[15]和 ECA-Net[16], 分别从通道或空间单一维度学习特征权重; 采用多路注意力的 SK-Net[17], ResNeSt[18]和 CBAM[19], 则以串行方式综合空间和通道 2 个分支通路的注意力信息;

采用双注意力的 DA-Net[20]和 PFAN[21]，以并行方式综合空间和通道 2 个分支通路的注意力信息. 近年来，注意力模型开始引起部分研究者的关注，并被用于自然语言处理、图像识别、语音识别等学习任务中. 此外，上述注意力模型利用特征的全局和局部信息提取高级别语义特征，以获得更准确的分割结果，但其忽略了网络模型中间层信息的特征表达能力.

为了解决上述问题，本文在 CNN 中引入注意力机制，使用特征融合有效地利用 CNN 中浅层蕴含的空间几何信息和深层丰富的语义信息，提高网络的分割精度. 通过注意力机制有选择地增强有效特征而抑制无效特征，在空间和通道 2 个维度上计算注意力分布，赋予不同尺度特征不同的权重来充分融合多尺度特征，高效地利用各尺度特征语义信息. 使用并行的不同扩张率的扩张卷积聚合多尺度目标信息，提升对不同尺度目标的分割效果. 针对 CNN 分割边界较为模糊、边界定位精度较低等问题，采用基于轮廓检测的深度学习方法对 CNN 分割边界进行细粒度提取，利用准确的内部对应像素语义信息替代边界特征，进一步提升网络分割精度. 实验结果表明，本文网络能够有效地提高复杂场景下多尺度目标的语义分割精度；与传统的 CNN 模型相比，在公共数据集 PASCAL VOC2012 和 Cityscapes 上取得了显著的分割效果.

## 1 本文网络

本文围绕图像语义分割中多尺度特征信息的提取和边界分割结果的优化，构建了联合注意力机制的深度学习网络 AttenAtrNet，包括注意力卷积、多尺度特征聚合及边界精细粒度特征提取 3 个模块，其架构如图 1 所示.

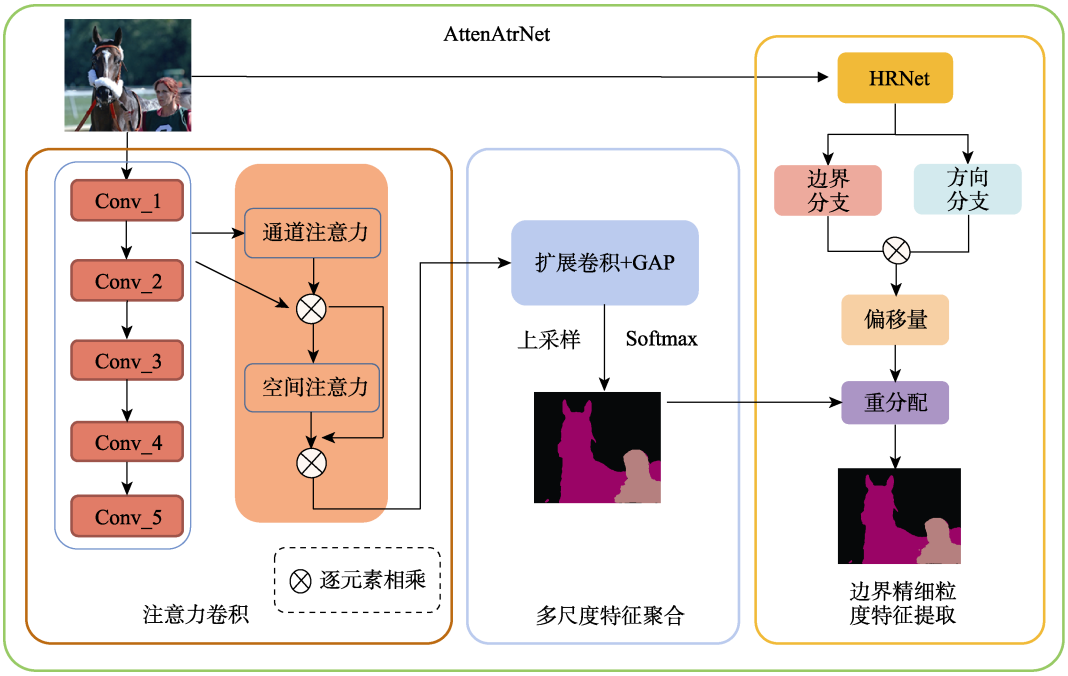



图 1 AttenAtrNet 架构

### 1.1 注意力卷积模块

本文使用跳跃映射的残差单元和扩张卷积作为骨干网络，称为 AttenAtrNet. 与普通网络的串行结构不同，残差单元中增加了跳跃映射，将输入与输出直接相加，补充卷积过程中损失的特征信息. AttenAtrNet 共包含 5 个卷积模块 Conv_1, Conv_2, Conv_3, Conv_4 和 Conv_5，分别包括 3, 9, 12, 69 和 9 个卷积层. 其中，Conv_4 和 Conv_5 块被设计为扩张卷积，可以在保持特征分辨率的前提下扩大感受野.

为了避免扩张卷积在扩大感受野时造成小尺度目标信息丢失，采用注意力机制对多尺度特征进行融合，以提取更加完整的空间信息和语义信息. 引入多路注意力机制[22]，对提取到的分辨率为原始图像 1/4, 1/8, 1/8 和 1/8 的中间层特征进行加权特征融合，得到含多尺度信息的特征图. 在特征融合时，设计 Concat 特征融合和 Add 特征融合 2 种不同的融合策略，如图 2a 和图 2b 所示. 其中，

Concat 特征融合是将 4 个不同尺度特征在通道维度上进行拼接，重点是对后 3 层 1/8 原图大小的特征图进行 2 倍上采样，保证各层特征分辨率大小一致后再进行 Concat 操作，然后通过 1×1 卷积获取融合后的多尺度特征；Add 特征融合采用级联模式，首先通过 1×1 卷积对最后一层特征通道数降维，确保其与第 3 层特征通道数一致，然后进行特征相加操作，得到融合了最后 2 层特征信息的新特征图，再使用同样的方法依次将新特征图与第 2 层特征、第 1 层特征融合，最终得到融合了 4 层特征信息的特征图. 通过大量实验对 2 种特征融合策略进行对比分析，验证了 Concat 操作的优势.

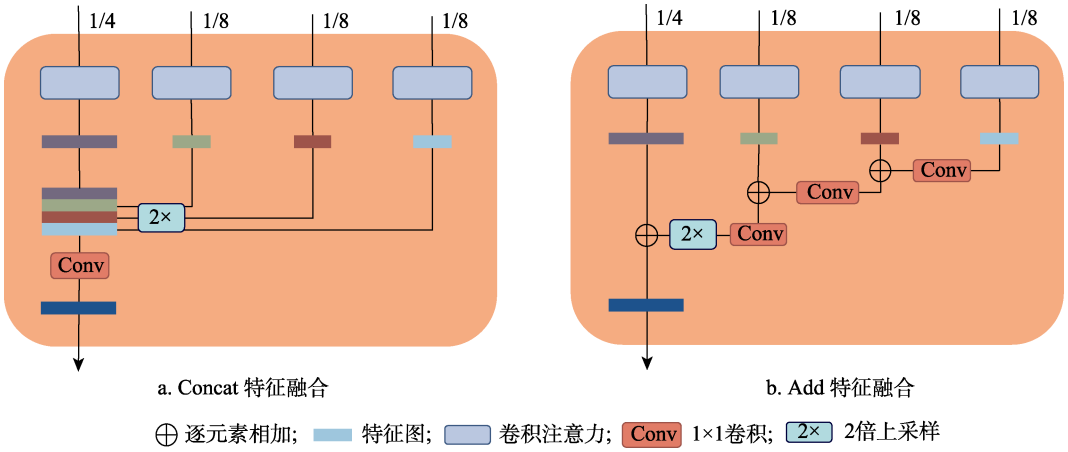



图 2　联合注意力机制的特征融合模块结构

卷积网络中的特征图不仅在通道中蕴含着丰富的注意力信息，通道内部(特征像素点之间)也含有大量的注意力信息[23]. 因此，本文采用基于多路注意力机制的卷积操作强化有效特征，抑制无效特征. 多路注意力包括通道注意力和空间注意力 2 个子模块. 输入的特征首先经过通道注意力子模块得到通道维度的注意映射，将注意映射与输入特征进行矩阵相乘；然后通过空间注意力子模块得到空间维度的注意映射；最后将得到的空间注意映射与通道注意力特征相乘，生成最终经过空间和通道 2 个维度优化的注意力特征.

通道注意力子模块利用通道间的特征关系生成通道注意力图，同时使用平均池化和最大池化对输入特征的空间维数进行压缩，能够有效地提高网络的表示能力. 该子模块获取通道注意力特征过程如下：首先使用平均池化和最大池化聚合特征图，生成 2 个通道特征 $F_{\text{avg}}^{\text{c}}$ 和 $F_{\text{max}}^{\text{c}}$；然后经过一个多层感知器(multi-layer perceptron, MLP)得到通道注意力特征

$$M_{\text{c}}(F)=\sigma\left(\text{MLP}\left(P_{\text{avg}}(F)\right)+\text{MLP}\left(P_{\text{max}}(F)\right)\right)=\sigma\left(W_1\left(W_0(F_{\text{avg}}^{\text{c}})\right)+W_1\left(W_0(F_{\text{max}}^{\text{c}})\right)\right) \tag{1}$$

其中，$\sigma$ 表示 Sigmoid 函数，MLP 的权值 $W_0$ 和 $W_1$ 对于 2 个输入是共享的；$P_{\text{avg}}$ 表示平均池化；$P_{\text{max}}$ 表示最大池化.

作为通道注意力的补充，空间注意力子模块利用特征的空间关系生成空间注意力图，关注“在哪里”的信息. 该子模块将通道注意力子模块得到的 $M_{\text{c}}(F)$ 作为输入特征，利用特征的空间关系生成空间注意力图. 首先沿着通道轴同时使用 $P_{\text{avg}}$ 和 $P_{\text{max}}$ 得到 2 个 2D 特征图 $F_{\text{avg}}^{\text{s}}$ 和 $F_{\text{max}}^{\text{s}}$；然后通过一个标准的卷积层将它们连接并卷积生成 2D 空间注意力图

$$M_{\text{s}}(F)=\sigma\left(f^{7\times7}\left[P_{\text{avg}}(F);P_{\text{max}}(F)\right]\right)=\sigma\left(f^{7\times7}\left[F_{\text{avg}}^{\text{s}};F_{\text{max}}^{\text{s}}\right]\right) \tag{2}$$

其中，$f^{7\times7}$ 表示卷积核大小为 7 的卷积操作.

## 1.2　多尺度特征聚合模块

复杂的现实场景图像中存在尺寸大小不一的目标，CNN 中重复的卷积池化操作会损失部分空间信息，导致小尺度目标的丢失；并且，传统卷积池化操作感受野大小固定，提取到的空间上下文信息有限，不利于多尺度目标的分割. 针对以上问题，考虑不同膨胀率的扩张卷积可以提取到不同尺度目标的信息，本文将输入的特征图并行通过不同膨胀率的扩张卷积和 GAP 以聚合多尺度目标信息，提出多尺度特征聚合模块；其结构如图 3 所示. 该模块共包含 1 个普通卷积层、3 个扩张卷积层和 1 个 GAP 层，其中，3 个卷积层的卷积核大小为 3×3，扩张率分别为 6，12 和 18，由注意力卷积

模块提取的注意力特征图首先以并行模式输入该模块的5个网络层，得到5个不同尺度的新特征图；然后将其在通道维度上进行聚合，通过1×1的卷积生成包含多尺度目标上下文信息的聚合特征.

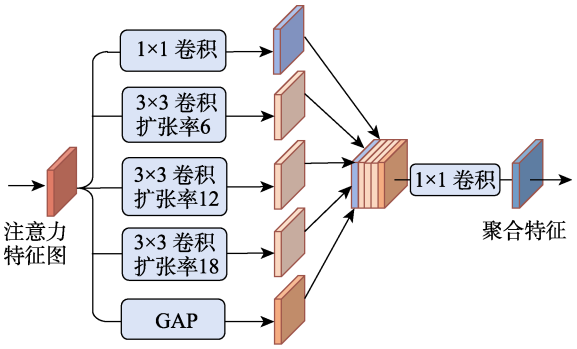



图 3 多尺度特征聚合模块结构

### 1.3 边界精细粒度特征提取模块

针对图像语义分割中因存在边界模糊，而影响整个分割精度的问题，受 SegFix[24]的启发，本文提出边界精细粒度特征提取模块. 以高分辨率网络 HRNet[13]为基本模块，学习从边界像素到内部像建立对应关系，用内部像素的可靠语义替代边界像素的不可靠预测信息，提高原有分割模型的分割准确性.

边界精细粒度特征提取的过程如图 4 所示.

(1) 以原始图像为输入，首先使用 HRNet 提取高分辨率图像特征，将其输入边界和方向2个平行分支；然后利用边界分支预测二值边界图，其中，0代表内部像素，1 代表边界像素；最后利用方向分支预测方向图，其中，每个元素存储从边界像素指向内部像素的方向，将方向范围划分为 8 个区域，对应的偏移量为 $(-1,1)$，$(0,1)$，$(1,1)$，$(1,0)$，$(1,-1)$，$(0,-1)$，$(-1,-1)$和$(-1,0)$，分别表示左上方、正上方、右上方、右方、右下方、正下方、左下方和左方，图 4 中用 8 个不同方向的箭头表示. (2) 将二值边界图与方向图逐元素相乘，保留边界部分的方向值，得到偏移图. (3) 利用偏移图对多尺度特征聚合模块的输出结果进行细化，即利用方向值指向的内部像素标签替代边界当前位置的语义信息，如图 4 中蓝色方框圈起来的像素点所示，利用偏移量重新分配类别标签. 考虑识别的边界会有误差，仅根据偏移量查找一个像素的距离可能并不能找到正确的内部像素，因此对所有的偏移量重新缩放一个因子，增加其查找的距离，确保找到正确的内部像素. 与 SegFix[24]不同，本文提出的边界精细粒度特征提取模块重新设置了缩放因子，动态地调整内部像素查找范围，确保以准确的内部对应像素语义替代边界语义特征，提高图像语义分割的准确性.

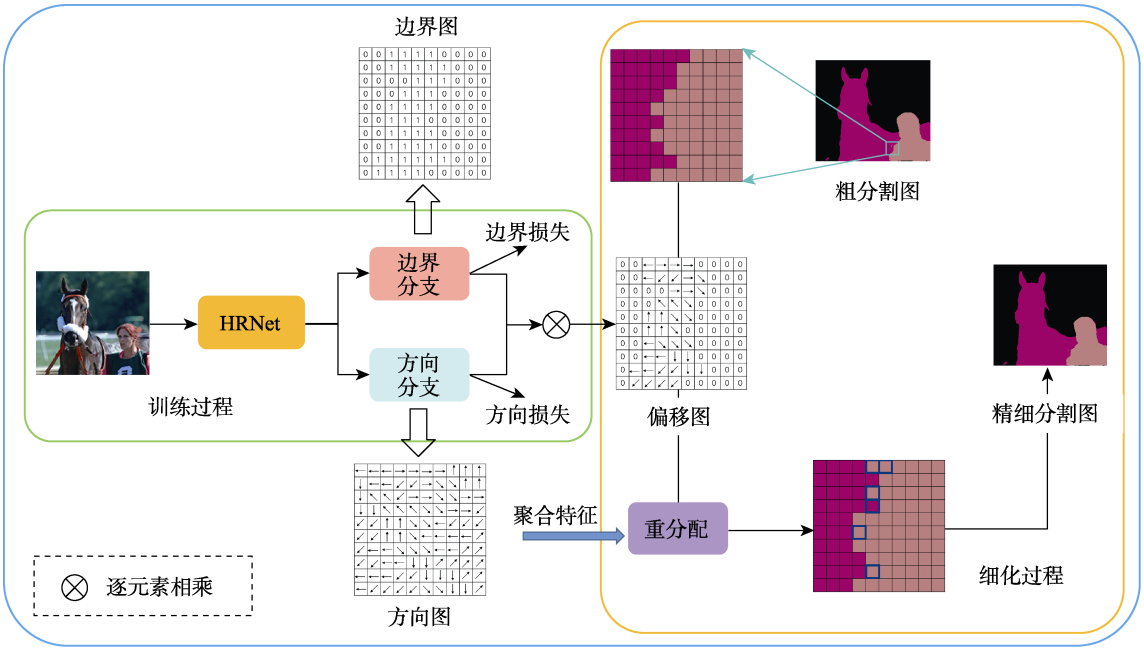



图 4 边界精细粒度特征提取模块

## 2 实验与结果分析

### 2.1 数据集和评价指标

为了验证本文构建的 AttenAtrNet 的性能，使用 PASCAL VOC2012 数据集[25](简称 VOC2012 数据集)和 Cityscapes 数据集[26]进行实验；并使用 SBD 数据集[27]扩充 VOC2012 数据集，得到训练集图像 10 582 幅，验证集 1 449 幅和测试集 1 456 幅. Ci-

tyscapes数据集包括50个不同城市的街道场景，采用 5000 幅精细标注图像中的 2975 幅用于训练，500幅用于验证，1525幅用于测试，共19个类别标注；每幅图像大小均为 2048×1024 像素，由于图幅中的道路场景复杂，因此目标类别尺度不一.

AttenAtrNet 在 Ubuntu18.04 系统上基于开源框架 PyTorch 实现，并使用 NVIDIA GeForce GTX 1080Ti 11GB GPU. 训练过程中，对图像随机缩小50%和放大 2 倍，并进行随机裁剪，防止在训练过程中出现过拟合. 为了增加训练样本的多样性，提高模型泛化能力，对输入图像进行−10°~10°随机旋转，并以 0.5 的概率对图像进行左右翻转；为了加快模型收敛速度，对输入图像进行标准化，减去各自所属数据集图像 RGB 的平均值并除以方差，使数据呈均匀分布状态. 对于 VOC2012 数据集，将图像裁剪为 321×321 像素，设置批处理大小为16，迭代 50 个 epoch；对于 Cityscapes 数据集，将输入图像裁剪为 713×713 像素，批处理大小设置为4，迭代100个epoch. 优化算法使用随机梯度下降(stochastic gradient descent, SGD)，动量设置为0.9，权重衰减系数设置为 0.0001，学习率使用poly 衰减策略，初始学习率设置为 0.01，学习率衰减系数设置为 0.9.

对于边界精细粒度特征提取模块，本文使用交叉验证确定其缩放因子. 将参数的搜索范围设置为 1~10，搜索步长设置为 1，分别在 VOC2012和 Cityscapes 验证集上搜索缩放因子的最佳值. 最终，在 VOC2012 数据集上缩放因子设置为 1，在Cityscapes 数据集上缩放因子设置为 3.

采用像素精度(pixel accuracy, PA)和平均交并比(mean intersection over union, mIoU)这 2 个指标评估 AttenAtrNet 的性能. 假设总共有 $k+1$ 个语义类(标记为 $L_0 \sim L_k$，包含一个背景类别)，其中，$P_{ij}$ 表示类别为 $i$ 的像素被预测为类别为 $j$ 的数目.

PA 表示预测正确的像素和总的像素的比率，公式为

$$\mathrm{PA}=\sum_{i=0}^{k}P_{ii}\Big/\sum_{i=0}^{k}\sum_{j=0}^{k}P_{ij} \tag{3}$$

mIoU 通过计算真实值集合和预测值集合的交集与并集之和的比值，计算图像真值与预测结果的重合程度. 该指标先基于每个类别计算，然后再求均值，公式为

$$\mathrm{mIoU}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}+\sum_{j=0}^{k}P_{ij}-P_{ii}} \tag{4}$$

## 2.2　对比实验

分别在 VOC2012 和 Cityscapes 数据集上对第1.1节中的2种特征融合策略进行实验，结果如表1所示. 可以看出，Concat 特征融合在 2 个数据集上的 MIoU 分别达到 77.5%和 70.8%，比 Add 特征融合的 76.5%和 69.8%均提高 1.0 个百分点，比未采用特征融合的原模型分别提高 12.0 个百分点和 7.7个百分点，充分论证了本文提出的特征融合策略的有效性，其中，Concat 特征融合策略对网络性能的提高更大，故本文模型最终选择 Concat 特征融合策略.

**表 1　2 种特征融合策略性能对比**

| 数据集 | 特征融合策略 | 评估指标/% | |
|---|---|---|---|
| | | PA | mIoU |
| VOC2012 | 原模型 | 91.0 | 65.5 |
| | Add 融合 | 94.5 | 76.5 |
| | Concat 融合 | **94.8** | **77.5** |
| Cityscapes | 原模型 | 93.0 | 63.1 |
| | Add 融合 | 94.4 | 69.8 |
| | Concat 融合 | **94.6** | **70.8** |

注. 粗体表示最优值.

在 VOC2012 验证集上，FCN-8S[5]，SegNet[7]，Deeplabv2[9]，PSPNet[11]和 Deeplabv3+[10]模型进行对比实验，验证本文构建的网络模型的性能，结果如表 2 所示. 可以看出，在相同的实验环境下，与其他模型相比，本文模型取得了最优的结果，mIoU 为 77.5%，比 Deeplabv2 提高 3.5 个百分点，比 PSPNet 提高 1.8 个百分点，比 Deeplabv3+提高1.3 个百分点，使用 SegFix 对边界进行优化后，mIoU 进一步提高 0.2 个百分点；FCN-8S 和 SegNet由于使用的主干网是 VGG-16，网络层数较浅，特征提取能力有限，分割效果较差，mIoU 分别为62.7%和 57.3%.

**表 2　在 VOC2012 数据集上 6 种模型性能对比**

| 模型 | 骨干网 | 评估指标/% | |
|---|---|---|---|
| | | PA | mIoU |
| FCN-8S[5] | VGG-16 | 90.7 | 62.7 |
| SegNet[7] | VGG-16 | 89.8 | 57.3 |
| Deeplabv2[9] | ResNet-101 | 94.0 | 74.0 |
| PSPNet[11] | ResNet-101 | 94.4 | 75.7 |
| Deeplabv3+[10] | Xception | 94.5 | 76.2 |
| 本文 | ResNet-101 | 94.8 | 77.5 |
| 本文(+边界细粒度特征提取) | ResNet-101 | **94.9** | **77.7** |

注. 粗体表示最优值.

表 3 所示为在 VOC2012 验证集上, 3 种模型对 21 个语义类别的 IoU 值对比. 可以看出, 在 21 个语义类别中, 有 14 个语义类别达到最优, 其他类别虽然没有达到最优, 但与最优值也较为接近, 说明本文提出的联合注意力机制的多尺度特征融合策略是有效的, 能够充分地利用不同尺度特征的有效信息, 减少网络由于重复卷积池化操作损失的语义信息, 提高网络的分割精度.

**表 3 在 VOC2012 数据集上 3 种模型对 21 个语义类别 IoU 值对比** %

| 类别 | PSPNet[11] | Deeplabv3+[10] | 本文 | 本文(+边界细粒度特征提取) | 类别 | PSPNet[11] | Deeplabv3+[10] | 本文 | 本文(+边界细粒度特征提取) |
|---|---|---|---|---|---|---|---|---|---|
| 背景 | 93.7 | 93.7 | 93.9 | **93.9** | 餐桌 | 40.9 | **51.4** | 47.9 | 48.1 |
| 飞机 | 89.7 | 90.2 | 89.5 | **90.3** | 狗 | 85.9 | 87.2 | 88.3 | **88.6** |
| 自行车 | 42.4 | 42.4 | 43.9 | **44.2** | 马 | **87.3** | 83.1 | 86.3 | 86.4 |
| 鸟 | **89.3** | 89.2 | 89.1 | 89.2 | 摩托车 | 84.7 | 86.6 | 87.4 | **87.8** |
| 船 | 70.2 | 68.8 | 70.6 | **71.0** | 人 | 84.9 | 84.6 | 85.3 | **85.4** |
| 瓶子 | 75.5 | 77.7 | 81.7 | **81.8** | 盆栽植物 | **61.7** | 59.0 | 61.4 | 61.6 |
| 公交车 | 93.1 | 93.5 | 94.1 | **94.1** | 羊 | 85.8 | 88.8 | 89.5 | **89.9** |
| 汽车 | 87.3 | **90.1** | 88.6 | 88.7 | 沙发 | **43.3** | 42.5 | 40.7 | 40.9 |
| 猫 | 89.5 | 91.4 | 92.2 | **92.4** | 火车 | 84.5 | 83.7 | 84.9 | **85.1** |
| 椅子 | 40.5 | 37.1 | 44.8 | **45.0** | 电视机 | 75.4 | 76.2 | **78.7** | 78.6 |
| 奶牛 | 85.0 | 84.1 | 87.7 | **87.9** | 平均 | 75.7 | 76.2 | 77.5 | **77.7** |

注. 粗体表示最优值.

在 Cityscapes 验证集上, 将本文模型与其他模型进行实验, 进一步验证该模型的泛化性, 结果如表 4 所示. 可以看出, 与其他模型相比, 本文模型取得了最好的分割效果, PA 为 94.6%; mIoU 为 70.8%, 比 PSPNet, Deeplabv3+分别提高 2.2 个百分点和 1.3 个百分点; 使用边界精细力度特征提取模块后, mIoU 进一步提高 1.6 个百分点.

在 Cityscapes 验证集上, 3 种模型对 19 个语义类别的 IoU 值对比如表 5 所示. 可以看出, 本文模型在 13 个语义类别中的分割精度达到最好, 其中, 骑手、卡车、火车等类别的分割效果远远高出其他模型, 说明 AttenAtrNet 能够充分地融合多尺度特征, 利用注意力机制关注有效信息, 忽略无效信息, 显著地提高网络的分割效果.

**表 4 在 Cityscapes 数据集上 6 种模型性能比较** %

| 模型 | PA | mIoU |
|---|---|---|
| FCN-8S[5] | 91.2 | 54.1 |
| SegNet[7] | 91.1 | 57.8 |
| Deeplabv2[9] | 92.6 | 63.8 |
| PSPNet[11] | 94.3 | 68.6 |
| Deeplabv3+[10] | 94.4 | 69.5 |
| 本文 | 94.6 | 70.8 |
| 本文(+边界细粒度特征提取) | **94.8** | **72.4** |

注. 粗体表示最优值.

**表 5 在 Cityscapes 数据集上 3 种模型对 19 个语义类别 IoU 值对比** %

| 类别 | PSPNet[11] | Deeplabv3+[10] | 本文 | 本文(+边界细粒度特征提取) | 类别 | PSPNet[11] | Deeplabv3+[10] | 本文 | 本文(+边界细粒度特征提取) |
|---|---|---|---|---|---|---|---|---|---|
| 马路 | 97.4 | **97.5** | 96.8 | 97.1 | 天空 | 93.1 | 93.1 | 93.6 | **94.2** |
| 人行道 | 80.6 | **81.6** | 77.1 | 79.0 | 人 | 77.0 | 77.3 | 77.2 | **80.6** |
| 建筑 | 90.2 | 90.3 | 90.0 | **90.7** | 骑手 | 50.6 | 52.7 | 56.2 | **58.4** |
| 围墙 | 45.0 | **45.1** | 35.9 | 36.5 | 汽车 | 93.2 | 93.4 | 92.0 | **93.5** |
| 栅栏 | 47.4 | 48.8 | 51.0 | **51.8** | 卡车 | 53.3 | 56.6 | 65.4 | **66.1** |
| 电线杆 | 58.2 | **58.4** | 53.2 | 56.8 | 公交车 | 64.4 | 65.9 | 68.9 | **69.8** |
| 红绿灯 | 60.6 | 61.6 | 63.5 | **66.0** | 火车 | 42.0 | 43.9 | 71.9 | **72.8** |
| 交通标志 | 71.1 | 71.6 | 71.4 | **75.2** | 摩托车 | 52.5 | 53.7 | 60.7 | **62.4** |
| 植物 | 91.5 | **92.1** | 90.8 | 91.5 | 自行车 | 72.2 | 73.5 | 73.8 | **75.9** |
| 绿化带 | 62.4 | **62.5** | 56.7 | 57.9 | 平均 | 68.6 | 69.5 | 70.8 | **72.4** |

注. 粗体表示最优值.

在相同的实验环境下，使用 10 幅分辨率大小为 513×513 像素的彩色图像对 6 种模型的推理速度进行实验，对比不同模型算法的效率，同时统计各种模型的可训练的参数量，结果如表 6 所示. 可以看出，本文模型有 61 334 365 个可训练参数，虽然少于 PSPNet，但其处理速度仅为 7.7 幅/s，在速度上不及 PSPNet，这是由于本文模型在多尺度特征聚合模块中采用扩张卷积，扩张卷积的计算量导致时间成本增加，这一点从 Deeplabv2 的计算速度上也可以得到印证；Deeplabv3+的参数量和速度都相对较好，这是因为其使用 Xception 作为骨干网，比本文模型采用的残差模块计算效率更高；AttenAtrNet 虽然取得了较好的分割结果，但其牺牲了一部分计算效率，从该角度分析，该网络的执行效率是未来优化的目标.

**表 6　不同模型的参数量及效率对比**

| 模型 | 参数量 | 处理速度/(幅·s$^{-1}$) |
|---|---|---|
| FCN-8S[5] | 134 362 751 | 50.0 |
| SegNet[7] | 49 834 687 | 25.0 |
| Deeplabv2[9] | 44 172 308 | 8.3 |
| PSPNet[11] | 65 708 501 | 11.1 |
| Deeplabv3+[10] | 54 705 317 | 20.0 |
| 本文 | 61 334 365 | 7.7 |

## 2.3　消融实验

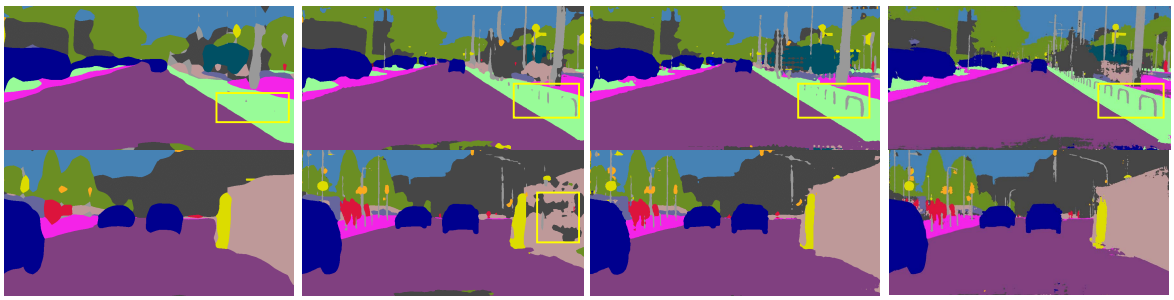
为了验证 AttenAtrNet 中各个模块的有效性，分别在 VOC2012 和 Cityscapes 数据集上进行消融实验，结果如表 7 所示. 可以看出，在 2 个数据集上，在残差模型的 Conv_4 和 Conv_5 块引入扩张卷积后，实验组 2 的 mIoU 比实验组 1 分别提高 6.6 个百分点和 3.2 个百分点，表明扩张卷积能够有效地避免大幅度下采样导致的语义信息丢失；加入注意力模块后，实验组 3 的 mIoU 比实验组 2 分别提高 4.8 个百分点和 2.0 个百分点，表明联合注意力机制的多尺度特征融合策略可以有效地利用不同尺度的细节特征，有选择地强化有效信息，忽略无效信息，提升模型的分割效果；加入多尺度特征聚合模块后，实验组 4 的 mIoU 比实验组 3 分别提高 0.6 个百分点和 2.5 个百分点，表明多尺度特征聚合模块可以高效地聚合不同尺度特征的空间上下文信息，提升模型的分割效果；加入边界精细粒度特征提取模块后，实验组 5 的 mIoU 比实验组 4 分别提高 0.2 个百分点和 1.6 个百分点，表明该边界优化模块能够进一步提高网络分割的准确性.

**表 7　在 2 个数据集上进行消融实验结果对比**

| 数据集 | 实验组 | 扩张卷积 | 注意力模块 | 多尺度聚合模块 | 边界精细粒度特征提取 | mIoU/% |
|---|---|---|---|---|---|---|
| VOC2012 | 1 | | | | | 65.5 |
| | 2 | √ | | | | 72.1 |
| | 3 | √ | √ | | | 76.9 |
| | 4 | √ | √ | √ | | 77.5 |
| | 5 | √ | √ | √ | √ | **77.7** |
| Cityscapes | 1 | | | | | 63.1 |
| | 2 | √ | | | | 66.3 |
| | 3 | √ | √ | | | 68.3 |
| | 4 | √ | √ | √ | | 70.8 |
| | 5 | √ | √ | √ | √ | **72.4** |

注. 粗体表示最优值.

对实验组 1~实验组 4 的分割结果进行可视化，以更加直观地展示各个模块的作用，如图 5 所示. 可以看出，仅使用基础模块时，得到的分割结果丢失了许多小目标细节信息，如第 1 行中黄框中的线杆未能被识别出来；在使用扩张卷积后的实验组 2 中，许多细节信息被保留下来，边界部分也更清晰，但第 2 行黄框部分的效果反而有所降低，出现了许多空洞；加入注意力机制后的实验组 3 中该问题得到解决，说明联合注意力机制的多尺度特征融合策略能够充分融合各尺度特征，丰富提取到的特征信息；加入多尺度特征聚合后的实验组 4，分割结果中小目标信息更加丰富并且分割也更加准确.

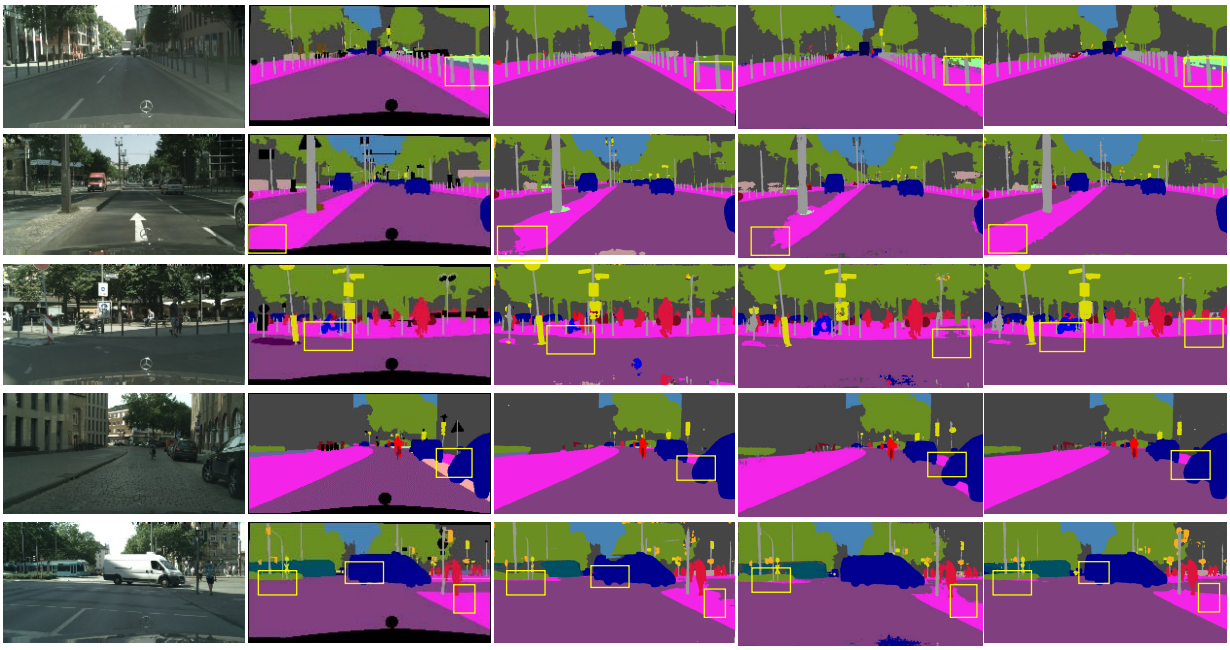


a. 实验组 1　　b. 实验组 2　　c. 实验组 3　　d. 实验组 4

图 5　各模块分割结果可视化

### 2.4 分割结果可视化

在 Cityscapes 数据集上将 AttenAtrNet 与 PSPNet 和 Deepabv3+的分割效果进行可视化, 以更直观地展示该网络的分割效果, 结果如图 6 所示. 可以看出, 第 1 行中, PSPNet 将绿化带像素误分为植物, Deeplabv3+也将绿化带的部分像素误分为植物, 而 AttenAtrNet 能够通过多尺度特征信息融合获取丰富的信息, 正确地分割出绿化带; 第 2 行中, PSPNet 和 Deeplabv3+没有将人行道像素正确地分割出来; 第 3 行中, PSPNet 将小部分摩托车像素误分成人, 第 4 行中, PSPNet 也未能分割出汽车的后视镜, Deeplabv3+分割出的人行道有部分缺失, 而 AttenAtrNet 对小尺度目标对象也有较好的分割结果; 第 3 行的黄框中, AttenAtrNet 的分割相对完整; 第 5 行中, 由于太阳光的照射, PSPNet 将汽车类部分像素预测为天空, 对火车类和人行道类也有误分, Deeplabv3+未能完全正确地分割, 而 AttenAtrNet 能够充分考虑空间上下文信息提取物体空间特征, 将其正确地分割出来. 实验结果表明, AttenAtrNet 能够充分利用多尺度信息提升分割效果, 并且对不同尺度的目标均有较好的分割结果.



a. 原图 b. 标签 c. PSPNet[11] d. Deeplabv3+[10] e. AttenAtrNet

图 6 Cityscapes 数据集分割结果可视化

## 3 结 语

本文针对传统深度 CNN 在进行语义分割时存在部分语义信息丢失、边界定位精度低、分割结果较为粗糙等问题, 利用 CNN 中间层特征具有丰富空间信息的特点, 构建一种联合注意力机制和多尺度特征的 CNN 模型. 以残差单元为基础模块, 使用扩张卷积保证输出特征的分辨率, 并联合注意力机制对不同层次特征进行加权融合, 丰富输出特征所含信息; 使用扩展卷积和 GAP 聚合多尺度目标信息, 增强网络的特征表达能力; 最后使用边界精细粒度特征提取模块优化分割边界. 实验结果表明, AttenAtrNet 在提取的语义信息的完整性和边界分割结果的准确性方面具有较好的性能, 可有效地对复杂自然场景图像进行 19 类目标的正确语义分割, 且性能优于现有的主流网络. 未来, 将进一步提高 AttenAtrNet 的提取效率.

## 参考文献(References):

[1] Zhang R, Li G Y, Li M L, *et al*. Fusion of images and point clouds for the semantic segmentation of large-scale 3D scenes based on deep learning[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 143: 85-96

[2] Fan Wei. Color image segmentation algorithm based on region growth[J]. Computer Engineering, 2010, 36(13): 192-193+196 (in Chinese)
(范伟. 基于区域生长的彩色图像分割算法[J]. 计算机工程, 2010, 36(13): 192-193+196)

[3] Zhang Qian, Feng Fujian, Lin Xin, *et al*. An image segmenta-

tion algorithm based on graph theory[J]. Computer Engineering, 2012, 38(18): 194-197(in Chinese)
(张乾，冯夫健，林鑫，等. 一种基于图论的图像分割算法[J]. 计算机工程, 2012, 38(18): 194-197)

[4] Yang Bisheng, Han Xu, Dong Zhen. A deep learning network for semantic labeling of large-scale urban point clouds[J]. Acta Geodaetica et Cartographica Sinica, 2021, 50(8): 1059-1067(in Chinese)
(杨必胜，韩旭，董震. 适用于城市场景大规模点云语义标识的深度学习网络[J]. 测绘学报, 2021, 50(8): 1059-1067)

[5] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440

[6] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] //Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241

[7] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495

[8] Chen L C, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs[OL]. [2022-10-19]. https://arxiv.org/abs/1412.7062

[9] Chen L C, Papandreou G, Kokkinos I, *et al*. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848

[10] Chen L C, Zhu Y K, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 833-851

[11] Zhao H S, Shi J P, Qi X J, *et al*. Pyramid scene parsing network[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6230-6239

[12] Lin G S, Milan A, Shen C H, *et al*. RefineNet: multi-path refinement networks for high-resolution semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 5168-5177

[13] Sun K, Zhao Y, Jiang B R, *et al*. High-resolution representations for labeling pixels and regions[OL]. [2022-10-19]. https://arxiv.org/abs/1904.04514

[14] Zhang Chenjia, Zhu Lei, Yu Lu. Review of attention mechanism in convolutional neural networks[J]. Computer Engineering and Applications, 2021, 57(20): 64-72(in Chinese)
(张宸嘉，朱磊，俞璐. 卷积神经网络中的注意力机制综述[J]. 计算机工程与应用, 2021, 57(20): 64-72)

[15] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7132-7141

[16] Wang Q L, Wu B G, Zhu P F, *et al*. ECA-Net: efficient channel attention for deep convolutional neural networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 11531-11539

[17] Li X, Wang W H, Hu X L, *et al*. Selective kernel networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 510-519

[18] Zhang H, Wu C R, Zhang Z Y, *et al*. ResNeSt: split-attention networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2022: 2735-2745

[19] Hu J, Shen L, Albanie S, *et al*. Gather-excite: exploiting feature context in convolutional neural networks[C] //Proceedings of the 32nd International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2018: 9423-9433

[20] Li Z Z, Qu N, Li X X, *et al*. Partial discharge detection of insulated conductors based on CNN-LSTM of attention mechanisms[J]. Journal of Power Electronics, 2021, 21(7): 1030-1040

[21] Liu Y F, Zhu Q G, Cao F, *et al*. High-resolution remote sensing image segmentation framework based on attention mechanism and adaptive weighting[J]. ISPRS International Journal of Geo-Information, 2021, 10(4): Article No.241

[22] Woo S, Park J, Lee J Y, *et al*. CBAM: convolutional block attention module[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-19

[23] Wang Y B, Wang H F, Peng Z H. Rice diseases detection and classification using attention based neural network and Bayesian optimization[J]. Expert Systems with Applications, 2021, 178: Article No.114770

[24] Yuan Y H, Xie J Y, Chen X L, *et al*. SegFix: model-agnostic boundary refinement for segmentation[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 489-506

[25] Everingham M, Eslami S M A, van Gool L, *et al*. The PASCAL visual object classes challenge: a retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136

[26] Cordts M, Omran M, Ramos S, *et al*. The cityscapes dataset for semantic urban scene understanding[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 3213-3223

[27] Hariharan B, Arbeláez P, Bourdev L, *et al*. Semantic contours from inverse detectors[C] //Proceedings of the International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 991-998